***BSc (Honours) in Computing in Software Development***

***Stage 3***

**Modules:** Big Data System, Data Analysis 2 **Assignment:** Final Project

**Due:** 19-12-19 **Credit:** BDS 35%, DA2 30%

**Objectives**

To practice the following:

* Data preparation for data mining
* Choosing appropriate databases to store and analyse the data
* Backing up the databases to make sure there is no loss in the event of a server failure
* Choosing databases that can easily scale to handle large amounts of data
* Aggregating data into similar time frames
* Analyse the data
* Carry out linear regression
* Carry out multiple linear regression
* Carry out cluster analysis and explain the clusters found
* Carry out principal component analysis

**The Data**

You are provided with four data files, hourly\_dublin\_17\_18, KeyHourly.txt, JLHome1718Power.csv and JLHome1718Temperature.csv. This is real data about the weather at Dublin Airport, and the temperature and energy use in a house in Dundalk.

The energy usage and temperature data is gathered using an [OpenEnergyMonitor](https://openenergymonitor.org/) [emonPi](https://wiki.openenergymonitor.org/index.php/EmonPi). OpenEnergyMonitor is a small company based in Wales that uses the emonPi to give real time energy usage monitoring.

**hourly\_dublin\_17\_18** contains weather data from Dublin Airport from 1-1-17 to 31-12-18. The weather data is recorded every hour.

**KeyHourly.txt** contains information about the columns that are in the hourly\_dublin\_17\_18 file.

**JLHome1718Power.csv** contains data from 12-10-17 to 13-10-19. This data measures the power usage in watts and each entry indicates the average usage over the previous minute.

**JLHome1718Temperature.csv** contains data from 12-10-17 to 13-10-19. This data measures the temperature of the house in degrees Celsius and each entry indicates the average temperature of the house over the previous 5 minutes.

**Requirements – Big Data Systems**

You are required to:

1. Take the data contained in the four files and store it in an appropriate database or databases.

When choosing your databases you should allow for the fact that there will be many houses being monitored using the emonPi and the ambition is to have every house in Britain and Ireland monitored. It is very important that your solution is able to scale to deal with billions of records and that it is backed up.

2. Detail how the databases you selected in part 1 will allow for the addition of potentially different structured data from different homes and weather stations.

3. Detail how your solution will scale when every home in Britain and Ireland wants to use an emonPi to monitor the home.

4. Detail how you will back up the data contained in the database and provide high availability to servers that need to query the data.

5. Detail how you will prepare, clean, aggregate and analyse the data. You should submit a list of database queries or a script which uses these queries and discuss them in a screencast.

6. Code repository. All code, documents and database dumps must be committed to a code repository and a link to this repository should be submitted as part of you submission on Moodle. Make sure that you have shared the repository with your lecturer before the submission date.

7. You will be required to give an **in-class demo** of your solution during class time **11:00-13:00 on Thursday December 19th.** This will be a joint presentation for Big Data Systems and Data Analysis 2.

8. Screencast. You are required to submit a screencast detailing your answers to parts 1 to 5 above. The screencast should be no longer than 15 minutes. It should be uploaded to YouTube and listed as unlisted or public. I suggest you use [OBS](https://obsproject.com/) to create the screencast.

**Requirements – Data Analysis 2**

You are required to explore the data for the **year 2018 only** and carry out the following:

1. Carry out an initial exploration of the entire data set. This should include, the identification and treatment of any outliers/missing data, the creation of appropriate visualisations and the generation of relevant descriptive statistics including correlation.
2. Carry out a k-means cluster analysis on the weather data. Justify your choice of k and detail how this choice was made. Investigate and describe the clusters.
3. Use principal component analysis to reduce the dimensionality in the weather dataset, identify the important principal components of your data set, interpret your results and create appropriate visualisations of your principal components.
4. Discuss potential research questions that can be analysed using this data set as a whole, where the response variable is the average Power usage in an hour.
5. Conduct a simple linear regression analysis: Determine the “best” simple linear regression model to predict the response variable. Justify your choice of model, check if the model has met the assumptions and interpret your results.
6. Conduct multiple linear regression analysis (analyses) to determine the most suitable model to predict the response variable. Justify your choice of model, check assumptions are met, interpret your results, and referring to the fit of the model. Compare this model to the simple linear regression analysis from part 5.
7. Put overall conclusion on all your results, summarizing your findings.
8. Submit your R code and a detailed report (with R output and visualisations) in which you detail your analysis and address questions 1-6 with all interpretations on Moodle by **23:55 Friday December 20th**
9. You will be required to give an **in-class demo** of your analysis during class time **11:00-13:00 on Thursday December 19th.** This will be a joint presentation for Big Data Systems and Data Analysis 2.

**Submission Requirements**

1. Source code, databases dumps, screencast urls and code repository urls and other items must be submitted in the relevant sub-folders inside a single ZIP file through Moodle.
2. Each student will be required to attend an **interview** after the deadline date. Each student will be questioned on the functionality of the code.
3. The assignment must be entirely the work of each student. Student are not permitted to share any pseudocode or source code from their solution with any other individual in the class. Students may not distribute the source code of their solution to any student in any format (i.e. electronic, verbal, or hardcopy transmission).
4. **Plagiarised assignments will receive a mark of zero**. This also applies to the individual allowing their work to be plagiarised.
5. Any plagiarism will be reported to the Head of Department and a report will be added to your permanent academic record.
6. Late assignments will only be accepted if accompanied by the appropriate medical note. This documentation must be received within 10 working days of the project deadline. The penalty for late submission is as follows:

•  Marked out of 80% if up to 24 hours late.

•  Marked out of 60% if 24-48 hours late.

•  Marked out of 40% if 48-72 hours late.

•  Marked out of 20% if 72-96 hours late.

•  Marked out of 0%, if over 96 hours late.